Cross-Dataset Transfer and Unknown-Class Detection in Imbalanced SAR Ship Classification

Ch Muhammad Awais[1 [0009−0001−4589−4103]], Marco Reggiannini[1 [0000−0002−4872−9541]], Davide Moroni[1 [0000−0002−5175−5126]], and Giulio Del Corso[1 [0000−0003−4604−2006]]

Institute of Science and Technology, National Research Council, Pisa, Italy
fname.lname@isti.cnr.it

Abstract. Ship classification from Synthetic Aperture Radar (SAR) imagery is a critical computer vision task, yet the robustness of models under deployment shifts remains unclear. While models are often trained on one dataset and deployed on another, we lack a comprehensive understanding of their cross-dataset generalization. To address this, we evaluate six pretrained models on two SAR ship datasets in three settings: in-domain classification, cross-dataset transfer, and unknown-class detection. For unknown detection, we hold out all classes one at a time. In-domain, SARDet100K gives the best balanced accuracy on both datasets (73.4% on FUSARShip and 53.2% on OpenSARShip). In cross-dataset transfer, we observe strong failures: some models show moderate accuracy but near-chance balanced accuracy (for example, 64.5% accuracy but 33.3% balanced accuracy for OpenSARShip to FUSARShip). In unknown detection, performance depends on the held-out class and dataset, while MC-dropout variance is often close to random. These findings show that cross-dataset generalization in SAR remains limited and that task-specific uncertainty scores are often more informative than MC-dropout variance for held-out-class detection, although their relative ranking depends on the dataset and held-out class.

Keywords: SAR Ship Classification · Uncertainty · imbalance

1 Introduction

Synthetic aperture radar (SAR) is a core sensing technology for Earth Observation (EO) because it operates in day and night conditions and can penetrate clouds, haze, and rain. This makes SAR especially important for maritime applications where optical imagery is often unreliable. SAR is used in ship traffic monitoring [20, 21], illegal fishing surveillance [9], border and coast guard operations [30, 35], search-and-rescue support, and early warning systems for maritime risk [29]. In these applications, reliable ship-type classification is not only a benchmark problem but also an operational requirement [2]. Classification errors can delay response actions or trigger false alarms in safety-critical workflows. For this reason, models must be evaluated not only for accuracy but also for reliability under changing operational conditions [3].

The methodology for SAR ship classification has progressed through clear stages. Early work relied on classical machine learning with hand-crafted features [6]. This was followed by deep learning models trained end-to-end for specific datasets. More recently, transfer learning with ImageNet-pretrained backbones became common with limited SAR-labeled data [2]. The current shift is toward domain-specific foundation models, which are trained at larger scale and are expected to provide better transferable representations [32, 34]. However, these models are still new in SAR, and their downstream behavior under realistic deployment shift is not yet well understood [1].

This gap motivates our study. We benchmark EO foundation-model [24] representations on SAR ship classification and evaluate not only predictive performance but also epistemic uncertainty under shift. This is important because models are often trained on one dataset and deployed on another with different sensors, acquisition geometry, preprocessing pipelines, and class balance. Under such changes, a model can retain moderate accuracy while failing on minority classes or unknown targets [5]. We therefore ask: how well do foundation-model features for SAR ship classification perform in-domain, across datasets, and when unknown classes appear?

To study this question, we introduce the SAR-Shift-Open Benchmark (SSOB), a comprehensive evaluation framework for six pretrained EO foundation models: DOFA, Prithvi, ScaleMAE, SSL4EO, SARDet100K, and SARJEPA. SSOB covers three settings: in-domain classification (ID), cross-dataset transfer (CD), and unknown-class detection (UC). In the unknown-class setting, we hold out Cargo, Fishing, and Tanker one at a time. We compare four uncertainty signals selected among the non-intrusive strategies that can be applied without systematic model modifications: *softmax confidence* [10] (commonly used even if often positively biased), *MC-dropout variance* [7] (easy extension at the cost of multiple forward passes), *class-oriented uncertainty*, and *domain-aware uncertainty* [8, 27] (more focused on anomaly detection).

Contributions.

1. We introduce SSOB, a comprehensive framework for evaluating pretrained representations on SAR ship classification under in-domain, cross-dataset, and held-out-class unknown settings.
2. We show that cross-dataset evaluation can hide failure when reported with accuracy alone, which motivates balanced metrics and per-class analysis.
3. We compare four post-hoc uncertainty scores for held-out-class detection and show that their behavior is strongly class and dataset dependent.

We do not propose a new backbone or training method. Our goal is to evaluate pretrained representations under a shared framework across matched, shifted, and held-out-class settings.

The paper is structured as follows: Section 2 introduces the SSOB evaluation framework and our selected uncertainty metrics. Section 3 details the experimental results across all three distribution shift settings, followed by a discussion of

operational implications and limitations in Section 4. Section 5 concludes the work.

2 Methodology

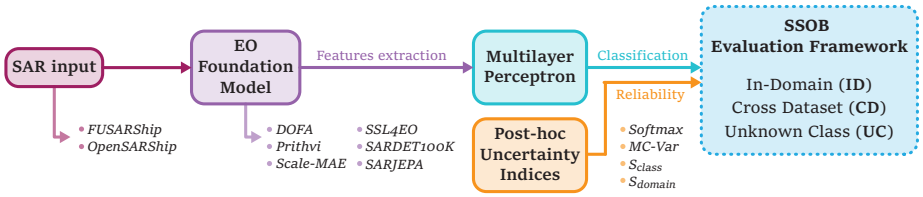

Fig. 1: Overview of the evaluation pipeline used in this study. SAR images are encoded with an Earth Observation (EO) foundation model to extract representations, followed by training of a shared classifier head. The trained model is then evaluated in three settings: in-domain, cross-dataset, and unknown-class settings. In-domain and cross-dataset branches report classification performance, while the unknown-class branch is used for uncertainty-based analysis.

Figure 1 summarizes the study pipeline. We evaluate SAR foundation-model representations by freezing each backbone and training a lightweight classifier head under a fixed setting.

2.1 Foundation Models

We benchmark six EO foundation models, DOFA [33], Prithvi [14], ScaleMAE [28], SSL4EO [31], SARDet100K [23], and SARJEPA [22], on SAR ship classification under a fixed training and evaluation setting. DOFA is a multimodal EO model designed to handle varying sensor types and channel configurations; Prithvi is a transformer-based geospatial model pretrained on HLS imagery; ScaleMAE learns scale-aware remote-sensing features via masked autoencoding; SSL4EO provides self-supervised weights trained on the SSL4EO-S12 Sentinel-1/Sentinel-2 corpus; SARDet100K is pretrained on large-scale SAR object detection; and SARJEPA is a SAR self-supervised model based on joint-embedding predictive learning.

2.2 Datasets

We evaluate all models on FUSARShip [11] and OpenSARShip [12], two public SAR ship datasets with different acquisition characteristics. FUSARShip is built from high-resolution GF-3 SAR images, while OpenSARShip is based on Sentinel-1 ship images. To ensure comparability and avoid extreme imbalance, we restrict

Table 1: Physical properties and instance counts; FUSARShip (FShip); OpenSARShip (OShip). 'Res.' = spatial resolution.

Dataset	Res.	Cargo	Tanker	Fishing
FShip	∼1.5 m	1693	785	148
OShip	∼20 m	5303	139	1825

both datasets to their shared ship categories (Cargo, Tanker, Fishing) and apply a consistent class mapping (Table 1). In all experiments, the backbone is frozen and only a lightweight classifier head is trained (described in the methodology section), so performance differences primarily reflect representation quality rather than end-to-end fine-tuning.

2.3 Classification Pipeline

Let a sample be (x, y), where x is a SAR image and $y \in \{\text{Cargo}, \text{Tanker}, \text{Fishing}\}$. For a pretrained model m, the backbone $\phi_m(\cdot)$ produces a feature vector $z = \phi_m(x)$, and a classifier head $h_\theta(\cdot)$ maps z to logits. The classifier is a multilayer perceptron consisting of a 512-dimensional hidden layer with ReLU nonlinearity, dropout probability 0.2 and a final linear projection to the class logits. The backbone is frozen all the time; only the classifier head is optimized.

Training uses cross-entropy loss, the Adam optimizer, learning rate 10^{-3}, weight decay 10^{-4}, and batch size 512 over up to 100 epochs. Early stopping monitors validation loss with patience 20 (minimum improvement 10^{-4}). Each configuration is evaluated over 5 data splits and 5 random seeds.

2.4 Uncertainty Scores

Let x denote an input SAR image. The frozen backbone ϕ_m maps x to a feature vector $z = \phi_m(x) \in \mathbb{R}^d$ with d ranging from 384 for SSL4EO to 2048 of SARDet100K backbones. Feature vectors are precomputed and cached during inference. A linear classifier head produces logits $a = Wz$ and class probabilities $p = \text{softmax}(a)$.

We evaluate four post-hoc uncertainty quantification methods:

Maximum softmax probability (Softmax). We use the complement of maximum softmax probability as a baseline uncertainty signal:

$$u_{\text{soft}}(x) = 1 - \max_k p_k(x).$$

This score reflects predictive confidence; higher values indicate lower model confidence, but it is known to be positively biased and tends to produce overconfident reliability predictions [5].

Classification Head MC-dropout variance (MC-Var) Monte Carlo dropout estimates epistemic (model) uncertainty by performing stochastic inference. With dropout enabled during inference, we execute $T = 10$ stochastic forward passes through the frozen backbone and classifier to make a fair comparison, producing predictions $\{p^{(t)}(x)\}_{t=1}^{T}$. The choice of using 10 forward passes is justified by [16]; while this small number of simulations may not guarantee perfect stability, it is sufficient to identify the most unreliable predictions. We compute class-wise variance and average across classes:

$$u_{\mathrm{mc}}(x) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{Var}_t \left(p_k^{(t)}(x) \right),$$

where $K = 3$ is the number of classes. Higher values indicate greater prediction variability across stochastic forward passes.

Class-oriented uncertainty (s_{class}). We compute class prototypes from the source training set as the mean feature representation per class:

$$\mu_c = \frac{1}{|D_c|} \sum_{(x_i, y_i) \in D_c} \phi_m(x_i), \qquad c \in \{1, 2, 3\},$$

where D_c is the subset of training samples with label c. Class-oriented uncertainty is defined as the minimum Euclidean distance to any class prototype:

$$s_{\mathrm{class}}(x) = \min_{c \in \{1,2,3\}} \|z - \mu_c\|_2 .$$

This score increases for samples that lie distant from all learned class clusters in feature space, indicating atypicality relative to known classes. Prototype-based scores are more reliable because they directly measure distance from the known training distribution in feature space, rather than deriving uncertainty from a classifier boundary that was never trained to represent out-of-distribution inputs. In addition, it has been proven [15] that they tend to converge to the optimal Bayesian approximator even without further re-training.

Domain-aware uncertainty (s_{domain}). From the same source training features, we compute a global domain center as the mean across all training samples:

$$\mu_{\mathrm{dom}} = \frac{1}{|D_{\mathrm{train}}|} \sum_{(x_i, y_i) \in D_{\mathrm{train}}} \phi_m(x_i).$$

Domain-aware uncertainty is the Euclidean distance to this center:

$$s_{\mathrm{domain}}(x) = \|z - \mu_{\mathrm{dom}}\|_2 .$$

This score increases as test samples deviate from the training-domain feature distribution, capturing domain shift.

2.5 Evaluation Framework

We evaluate model performance across three distinct settings: in-domain (ID), cross-dataset (CD), and unknown-class (UC).

In-Domain Evaluation Models are trained and tested on the same dataset using five-fold cross-validation. This setting measures supervised performance under distribution match.

Cross-Dataset Evaluation We evaluate transfer learning by training on one dataset and testing on the other, reporting both directions (FUSARShip to OpenSARShip and OpenSARShip to FUSARShip). This setting measures robustness to dataset shift while maintaining known-class structure.

Unknown-Class Evaluation To simulate out-of-distribution detection, we conduct unknown-class experiments by holding out each class during training. For each class $c \in \{\text{Cargo, Tanker, Fishing}\}$, all samples with label c are removed from the training set, and the classifier head is trained on the two remaining known classes. At test time, evaluation is performed on the full test set with class c treated as unknown. Uncertainty scores are computed for all test samples, and unknown-detection performance is quantified via AUROC to measure the ability of each score to distinguish held-out samples from known classes. This procedure is repeated three times to provide independent results for each held-out class.

3 Results

For classification, we report accuracy, balanced accuracy, macro-F1, and per-class recall. Balanced accuracy and macro-F1 serve as primary metrics under class imbalance. For uncertainty evaluation, we report the AUROC (Area Under the Receiver Operating Characteristic curve) of the uncertainty metrics against the agreement between model predictions and ground truth. This is a threshold-free metric bounded in [0,100%], where the maximum value corresponds to an uncertainty score that perfectly identifies mismatched cases (i.e., model prediction $\neq$ ground truth), 50% represents random performance (i.e., no discriminative ability), and values in [0,50%] correspond to indices that tend to assign higher uncertainty to correct predictions rather than to misclassified ones.

3.1 In-domain classification

In-domain results are shown in Table 2. SARDet100K is the top model on both datasets by balanced accuracy. The performance gap is substantial on FUSARShip (73.42% compared to the second-highest, SARJEPA at 58.34%) and remains distinct on OpenSARShip (53.22% vs. 45.60%). This supports the benefit of SAR-specific, task-aligned pretraining for in-domain robustness.

Table 2: In-domain results. Best value per metric is **bold**. Abbreviations: SD=SARDet100K, DF=DOFA, SJ=SARJEPA, S4=SSL4EO, SM=ScaleMAE, PV=Prithvi, Acc=Accuracy, BAcc=Balanced Accuracy.

	FUSARShip			OpenSARShip		
Model	Acc	BAcc	Macro-F1	Acc	BAcc	Macro-F1
SD	**75.68±1.97**	**73.42±3.98**	**75.40±3.16**	**76.05±0.58**	**53.22±3.38**	**56.66±3.37**
DF	69.58±1.63	56.57±2.68	59.32±2.06	74.62±0.54	44.38±2.62	47.09±3.41
SJ	69.20±2.28	58.34±3.49	61.26±3.30	74.36±0.43	45.60±2.66	48.31±2.96
S4	68.94±1.65	57.51±2.33	60.38±2.41	73.71±0.49	38.17±1.51	37.62±2.62
SM	66.49±1.66	50.91±4.75	53.53±5.12	73.16±0.21	38.69±2.09	36.99±3.17
PV	66.33±1.81	47.21±3.93	49.58±5.03	73.12±0.06	35.95±1.01	32.92±1.76

Table 3: Cross-dataset results. Best value per metric is **bold**. Rows are ordered by F→O rank; O→F values are reordered to match.

	FUSARShip → OpenSARShip			OpenSARShip → FUSARShip		
Model	Acc	BAcc	Macro-F1	Acc	BAcc	Macro-F1
SJ	**58.15±6.50**	29.25±2.04	27.88±1.15	57.11±2.95	30.46±1.25	25.18±0.77
DF	56.48±4.64	**45.63±4.53**	29.55±1.90	60.79±1.08	31.58±0.63	25.77±0.74
SM	54.31±2.32	30.71±0.67	30.42±0.74	64.11±0.77	33.10±0.40	26.06±0.19
PV	51.13±6.35	43.40±5.09	27.94±1.53	**64.49±0.32**	**33.34±0.16**	**26.30±0.36**
SD	48.01±1.77	36.81±0.74	**31.77±0.99**	53.63±3.98	31.98±4.07	25.97±2.24
S4	29.42±2.03	39.71±5.70	21.19±1.55	63.82±0.37	32.95±0.18	25.98±0.11

3.2 Cross-dataset transfer

Cross-dataset results in Table 3 show strong degradation from in-domain performance and clear direction asymmetry. In OpenSARShip to FUSARShip transfer, several models have moderate accuracy but near-chance balanced accuracy. The clearest example is Prithvi with 64.49 accuracy and 33.34 balanced accuracy. This is consistent with majority-class prediction and weak minority recall under shift. In FUSARShip to OpenSARShip transfer, the best balanced accuracy is higher (DOFA, 45.63) but remains far below in-domain values.

3.3 Unknown-class detection

Table 4 summarizes known-only classification performance (Acc, BalAcc, and Macro-F1) together with unknown-class detection AUROC. Across all combinations, the SARDet100K model achieves the strongest known-only classification.

Performance is consistently higher on FUSARShip, peaking when Cargo is held out (97.17 Acc, 92.60 BalAcc, 94.53 Macro-F1), whereas OpenSARShip is more challenging, with the weakest known-only results when Fishing is held out (77.09 Acc, 61.21 BalAcc, 62.28 Macro-F1).

Table 4: Unknown-class AUROC. For each dataset and held-out class, each column reports the best model value for that score.

Unknown	Acc (SD)	BAcc (SD)	Mac-F1 (SD)	Softmax	MC-Var	s_{class}	s_{domain}
FUSARShip							
Cargo	97.17 ±0.92	92.60 ±1.63	94.53 ±1.64	62.87 (PV)	51.72 (PV)	59.58 (SD)	55.35 (SD)
Fishing	97.68 ±0.33	88.65 ±3.03	91.59 ±1.40	42.06 (SD)	49.02 (SD)	66.29 (SJ)	58.05 (SM)
Tanker	*75.87 ±1.72*	*70.49 ±2.73*	*71.03 ±2.54*	42.51 (DF)	50.17 (DF)	**87.94** (PV)	**86.19** (PV)
OpenSARShip							
Cargo	95.21 ±0.66	74.27 ±4.56	78.38 ±3.96	45.83 (SJ)	49.76 (SJ)	57.91 (DF)	55.55 (DF)
Fishing	*77.09 ±0.78*	*61.21 ±2.16*	*62.28 ±2.83*	74.98 (PV)	51.96 (SJ)	**82.88** (SM)	**83.67** (SM)
Tanker	91.98 ±0.28	67.81 ±5.94	72.86 ±6.04	58.36 (SD)	51.50 (SD)	48.00 (SM)	48.67 (SM)

Beyond basic classification, unknown-class detection results show strong class and dataset dependence. For Fishing on OpenSARShip, ScaleMAE is strongest with 82.88 AUROC for class-oriented uncertainty and 83.67 for domain-aware uncertainty. For Tanker on FUSARShip, Prithvi is strongest with 87.94 (class-oriented) and 86.19 (domain-aware). For Cargo on OpenSARShip, the best values are 57.91 (class-oriented, DOFA) and 55.55 (domain-aware, DOFA). In contrast, Tanker on OpenSARShip is difficult for class-oriented and domain-aware scores, with best values near 48 to 49.

Across held-out classes, MC-dropout variance remains near random in most settings (about 49 to 52 AUROC). Softmax uncertainty is unstable: it is strong in some cases (for example, Fishing on OpenSARShip with 74.98) and weak in others (for example, Tanker on FUSARShip, where the best Softmax AUROC is only 42.51). By contrast, class-oriented and domain-aware uncertainty are the two strongest methods overall, and their best values in the 80–90 AUROC range are particularly high and noteworthy for unknown-class detection in SAR.

4 Discussion

The results show that model quality depends strongly on evaluation regime. In-domain performance and cross-dataset performance are not aligned, and unknown detection quality is not aligned with either of them. This has an important implication: a single leaderboard based on one metric or one setting is not enough to characterize reliability for SAR deployment. Per-class analysis (Appendix 6, Tables 8 and 10) reveals that moderate aggregate accuracy can mask catastrophic failure on individual classes under shift (e.g., Prithvi OpenSARShip to FUSARShip: 99.76% Cargo but 0.00% Tanker/Fishing), demonstrating why uncertainty-based error detection is essential when class-specific reliability cannot be guaranteed. Furthermore, while distance-based metrics (s_{class} and s_{domain}) excel at unknown-class detection, Supplementary Tables 7 and 9 show that they are much less effective for standard error detection within known classes, often remaining near 0.50 AUROC in ID and CD. This suggests that these metrics are primarily sensitive to domain and feature shift rather than mere classification ambiguity.

In-domain, SARDet100K is consistently the strongest model by balanced accuracy on both datasets. This suggests that SAR-specific pretraining improves representation quality when train and test distributions are matched. However, this advantage does not transfer directly to cross-dataset settings. In cross-dataset transfer, all models degrade substantially, and some configurations collapse to majority-class behavior. The clearest example is OpenSARShip to FUSARShip with Prithvi, where accuracy remains moderate (64.49%) but balanced accuracy drops to near chance (33.34%). This gap confirms that accuracy can hide severe class-wise failure under shift.

The direction asymmetry in cross-dataset transfer is also meaningful. FUSARShip to OpenSARShip and OpenSARShip to FUSARShip produce different rankings and different failure profiles, even under the same training settings. This indicates that domain shift is not a scalar property; it depends on direction and likely reflects differences in class priors, sensor characteristics, and feature statistics between datasets. For practical benchmarking, this means one-way transfer reporting is insufficient and can misstate deployment risk.

Unknown-class results further show that open-set behavior is class-dependent, dataset-dependent, and distinctly decoupled from in-distribution classification performance. For example, holding out Tanker on FUSARShip yields the lowest BalAcc for that dataset (70.49) but the highest unknown-class detection score (87.94). Conversely, holding out Cargo on FUSARShip produces the highest BalAcc (92.60) but weak open-set detection. This disconnect likely reflects differences in feature-space geometry across datasets. When the known classes are well separated from each other but the unknown class occupies a region of feature space close to the known distribution, distance-based scores lose discriminative power regardless of baseline classification accuracy.

The uncertainty comparison shows a consistent trend. MC-dropout variance is often near random ranking across unknown settings, which is expected here because dropout is applied only in the shallow classification head over a frozen

Table 5: **Practical guidance for evaluation.** Recommended metrics, uncertainty use, and decision rules across operating regimes.

Scenario	Report	Uncertainty
In-domain	BA, macro-F1	Secondary
Cross-dataset	BA + class recall; accuracy secondary	Required for screening
Unknown-class	AUROC; multiple held-out classes	Required for reject/flag
Pre-deployment policy	Bidirectional CD + multi-class UC	Use multiple scores, not only softmax/MC-dropout

backbone, so it captures only head-level stochasticity and remains largely insensitive to distributional shift in the input features, providing poor uncertainty measures as expected from existing literature [4, 26]. Softmax uncertainty is unstable, as expected from a score that is intrinsically positively biased [18]. By contrast, class-oriented and domain-aware scores are the strongest methods among those considered and are generally the most informative for unknown detection in this setting. As a further remark, the selected uncertainty measures are capable of identifying only a component of the full predictive uncertainty (i.e., epistemic uncertainty), whereas aleatoric uncertainty requires different approaches based on more intrusive methodologies [13].

These findings translate into a practical evaluation policy. In shifted settings, balanced accuracy and macro-F1 should be reported first, with accuracy kept as a secondary metric [3]. Based on these results, prototype-based scores (whether class-oriented or domain-aware) should be the primary uncertainty signal in open-set SAR evaluation [19, 25]. Softmax confidence and MC-dropout variance may be reported for completeness but should not be used as the sole basis for reject/flag decisions. Table 5 summarizes the recommended metrics, and the use of uncertainty.

This study has limits. It uses two datasets, three classes, and frozen-backbone training with a shallow nonlinear classifier head. These choices improve comparability but limit coverage of broader SAR conditions and adaptation strategies. The work also focuses on ranking metrics (AUROC) rather than threshold-calibrated decision analysis, which is necessary for operational deployment policies. In addition, while repeated runs improve statistical stability, the present results do not yet include external sensor families beyond the two benchmark datasets.

Several next steps follow directly from these limitations. Adding more datasets and sensor domains would test whether the observed direction asymmetry and class-dependent unknown behavior persist. Evaluating calibration-aware operating points and cost-sensitive thresholds would connect ranking performance to actionable decision rules. Similarly, intrusive Bayesian methods and multi-heads neural networks can provide better calibrated uncertainty estimates at the cost

of higher training cost [17], [5]. Finally, combining representation learning with explicit domain adaptation or calibration objectives may reduce cross-dataset collapse while preserving known-class performance.

5 Conclusion

This work benchmarks six foundation-model representations for SAR ship classification under in-domain evaluation, cross-dataset transfer, and unknown-class detection. The results show a clear split between in-domain and shifted behavior: SARDet100K is strongest in-domain by balanced accuracy on both datasets, while cross-dataset transfer remains fragile and can collapse to majority-class prediction, where moderate accuracy coexists with near-chance balanced accuracy. Unknown detection results also show that robustness is class-dependent and dataset-dependent, so conclusions from a single held-out class are incomplete. Across held-out classes, class-oriented and domain-aware uncertainty scores are generally more informative than MC-dropout variance and internal score, such as Softmax values, which are often near random. Overall, the main contribution is a reproducible evaluation framework that reveals failure modes that are hidden by accuracy-only reporting, and the practical recommendation is to use balanced metrics, bidirectional transfer tests, and multi-class unknown evaluation as standard components of SAR model assessment under distribution shift.

References

1. Awais, C.M., Reggiannini, M., Moroni, D., Karakus, O.: Fusion of foundation models for imbalanced sar ship classification. In: 4th ICLR Workshop on Machine Learning for Remote Sensing (Main Track) (2026)
2. Awais, C.M., Reggiannini, M., Moroni, D., Salerno, E.: A survey on sar ship classification using deep learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing pp. 1–26 (2026). https://doi.org/10.1109/JSTARS.2026.3695704
3. Awais, C., Reggiannini, M.: Deep learning for SAR ship classification: Focus on unbalanced datasets and inter-dataset generalization. In: 2024 International Conference on Electromagnetics in Advanced Applications (ICEAA). pp. 1–1 (2024). https://doi.org/10.1109/ICEAA61917.2024.10701968
4. Chan, A., Alaa, A., Qian, Z., Van Der Schaar, M.: Unlabelled data improves bayesian uncertainty calibration under covariate shift. In: International conference on machine learning. pp. 1392–1402. PMLR (2020)
5. Del Corso, G., Colantonio, S., Caudai, C.: Shedding light on uncertainties in machine learning: formal derivation and optimal model selection. Journal of the Franklin Institute **362**(3), 107548 (2025)
6. Feng, X., Zheng, H., Hu, Z., Yang, L., Zheng, M.: Dual-stream contrastive predictive network with joint handcrafted feature view for SAR ship classification. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7810–7814 (2024). https://doi.org/10.1109/ICASSP48485.2024.10446741

7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
8. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3614–3631 (2020)
9. Guan, Y., Zhang, X., Chen, S., Liu, G., Jia, Y., Zhang, Y., Gao, G., Zhang, J., Li, Z., Cao, C.: Fishing vessel classification in SAR images using a novel deep learning model. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–21 (2023). https://doi.org/10.1109/TGRS.2023.3312766
10. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
11. Hou, X., Ao, W., Song, Q., Lai, J., Wang, H., Xu, F.: Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. Science China Information Sciences **63**(4), 140303 (2020)
12. Huang, L., Liu, B., Li, B., Guo, W., Yu, W., Zhang, Z., Yu, W.: Opensarship: A dataset dedicated to sentinel-1 ship interpretation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **11**(1), 195–208 (2017)
13. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning **110**, 457 – 506 (2019), https://api.semanticscholar.org/CorpusID:216465307
14. Jakubik, J., Roy, S., Phillips, C., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., et al.: Foundation models for generalist geospatial artificial intelligence. arXiv preprint arXiv:2310.18660 (2023)
15. Jiang, H., Kim, B., Guan, M., Gupta, M.: To trust or not to trust a classifier. Advances in neural information processing systems **31** (2018)
16. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
17. Kristiadi, A., Hein, M., Hennig, P.: Being bayesian, even just a bit, fixes overconfidence in relu networks. In: International conference on machine learning. pp. 5436–5446. PMLR (2020)
18. Kuleshov, V., Liang, P.S.: Calibrated structured prediction. Advances in Neural Information Processing Systems **28** (2015)
19. Li, H., Song, J., Gao, L., Zhu, X., Shen, H.T.: Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. In: NeurIPS (2023)
20. Li, J., Xu, C., Su, H., Gao, L., Wang, T.: Deep learning for SAR ship detection: Past, present and future. Remote Sensing **14**(11), 2712 (2022)
21. Li, J., Yu, Z., Yu, L., Cheng, P., Chen, J., Chi, C.: A comprehensive survey on sar atr in deep-learning era. Remote Sensing **15**(5), 1454 (2023)
22. Li, W., Yang, W., Liu, T., Hou, Y., Li, Y., Liu, Z., Liu, Y., Liu, L.: Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. ISPRS Journal of Photogrammetry and Remote Sensing **218**, 326–338 (2024)
23. Li, Y., Li, X., Li, W., Hou, Q., Liu, L., Cheng, M.M., Yang, J.: Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. Advances in Neural Information Processing Systems **37**, 128430–128461 (2024)
24. Lu, S., Guo, J., Zimmer-Dauphinee, J.R., Nieusma, J.M., Wang, X., VanValkenburgh, P., Wernke, S.A., Huo, Y.: Vision foundation models in remote sensing: A survey. IEEE Geoscience and Remote Sensing Magazine **13**(3), 190–215 (2025)

25. Ma, X., Ji, K., Feng, S., Zhang, L., Xiong, B., Kuang, G.: Open set recognition with incremental learning for sar target classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023). https://doi.org/10.1109/TGRS.2023.3283423
26. Magris, M., Iosifidis, A.: Bayesian learning for neural networks: an algorithmic survey. Artificial Intelligence Review **56**(10), 11773–11823 (2023)
27. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems **32** (2019)
28. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4088–4099 (2023)
29. Reggiannini, M., Righi, M., Tampucci, M., Lo Duca, A., Bacciu, C., Bedini, L., D'Errico, A., Di Paola, C., Marchetti, A., Martinelli, M., Mercurio, C., Salerno, E., Zizi, B.: Remote sensing for maritime prompt monitoring. Journal of Marine Science and Engineering **7**(7) (2019). https://doi.org/10.3390/jmse7070202, https://www.mdpi.com/2077-1312/7/7/202
30. Topouzelis, K.N.: Oil spill detection by SAR images: Dark formation detection, feature extraction and classification algorithms. Sensors **8**(10), 6642–6659 (2008)
31. Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X.: Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. IEEE Geoscience and Remote Sensing Magazine **11**(3), 98–106 (2023)
32. Xiao, A., Xuan, W., Wang, J., Huang, J., Tao, D., Lu, S., Yokoya, N.: Foundation models for remote sensing and earth observation: A survey. IEEE Geoscience and Remote Sensing Magazine (2025)
33. Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Le Saux, B., Camps-Valls, G., Zhu, X.X.: Neural plasticity-inspired foundation model for observing the earth crossing modalities. CoRR (2024)
34. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 558–567 (2021)
35. Zilman, G., Zapolski, A., Marom, M.: The speed and beam of a ship from its wake's SAR images. IEEE Transactions on Geoscience and Remote Sensing **42**(10), 2335–2343 (2004)

6 Supplementary Results

Notation

Abbreviations used throughout: SD=SARDet100K, DF=DOFA, SJ=SARJEPA, S4=SSL4EO, SM=ScaleMAE, PV=Prithvi. Best values per dataset/direction are bold.

Table 6: Unknown-class AUPR (complement to main Table 4). Best model per held-out class.

Data	Unk.	Smax	MC-V	s_{class}	s_{domain}
FS	Cargo	**74.17**±2.59 (PV)	66.08±2.88 (PV)	73.30±2.21 (SD)	69.87±2.04 (SD)
FS	Fishing	24.99±1.32 (SD)	30.73±2.69 (S4)	46.13±3.86 (SJ)	40.94±3.47 (SJ)
FS	Tanker	5.00±0.34 (DF)	9.61±4.17 (PV)	40.42±4.16 (SM)	37.94±4.84 (SM)
OS	Cargo	69.83±1.73 (S4)	**73.06**±0.80 (S4)	78.39±0.56 (S4)	76.62±0.51 (S4)
OS	Fishing	6.55±3.77 (PV)	2.28±0.36 (SJ)	23.17±7.40 (SM)	23.46±7.35 (SM)
OS	Tanker	29.69±1.99 (SD)	26.42±1.28 (SD)	24.77±0.72 (SD)	24.96±1.05 (SD)

Table 7: Error detection AUROC: in-domain. No single uncertainty score is consistently strongest across all models and datasets; MC-dropout remains near random, while prototype-based scores vary substantially by model and dataset.

Model	s_{class}	s_{domain}	MC-Var	Bal. Acc
FUSARShip				
SD	42.49 ± 2.85	41.49 ± 2.13	51.11 ± 4.05	73.42 ± 3.98
SJ	**56.94 ± 2.63**	**55.41 ± 3.15**	50.95 ± 3.49	58.34 ± 3.49
S4	52.16 ± 3.46	51.16 ± 3.28	50.78 ± 3.22	57.51 ± 2.33
DF	47.61 ± 2.70	50.11 ± 2.24	50.54 ± 3.40	56.57 ± 2.68
SM	56.10 ± 4.05	56.52 ± 2.32	51.20 ± 4.44	50.91 ± 4.75
PV	47.65 ± 1.75	51.92 ± 4.29	50.90 ± 3.66	47.21 ± 3.93
OpenSARShip				
SD	47.29 ± 1.72	49.94 ± 1.93	51.89 ± 1.46	53.22 ± 3.38
SJ	42.67 ± 1.14	47.05 ± 1.50	51.35 ± 2.02	45.60 ± 2.66
DF	43.29 ± 1.35	47.10 ± 1.25	51.22 ± 1.75	44.38 ± 2.62
SM	**49.29 ± 1.18**	**51.02 ± 1.48**	52.18 ± 1.56	38.69 ± 2.09
S4	43.56 ± 0.93	47.36 ± 1.11	52.09 ± 1.58	38.17 ± 1.51
PV	45.33 ± 0.69	48.12 ± 1.49	51.85 ± 1.94	35.95 ± 1.01

Table 8: In-domain per-class accuracy. Columns show Cargo, Tanker, Fishing recognition rates. Imbalance in class performance signals potential brittleness under shift.

Model	**FUSARShip**			**OpenSARShip**		
	C	T	F	C	T	F
SD	**85.54±3.37**	**81.60±7.84**	**53.13±8.48**	**93.12±1.82**	29.40±5.44	37.14±9.90
SJ	85.49±4.68	52.53±11.05	37.01±9.66	96.53±1.28	13.56±3.27	26.71±8.03
S4	89.12±3.47	55.73±6.50	27.67±6.92	98.50±0.56	6.56±1.92	9.43±4.03
DF	87.82±2.62	47.73±8.15	34.17±6.43	96.79±0.83	14.20±2.88	22.14±6.93
SM	87.91±4.25	39.73±12.99	25.10±11.00	99.72±0.21	0.35±0.46	16.00±6.24
PV	90.32±4.91	30.13±9.45	21.17±9.67	100.00±0.02	0.00±0.00	7.86±3.03

Table 9: Error detection AUROC: cross-dataset. Prototype scores (35–79) capture distribution shift; MC-dropout ineffective. Example: Prithvi achieves 75.60 s_{class} despite 43.40% balanced accuracy, flagging errors where accuracy alone fails.

Direction	Model	s_{class}	s_{domain}	MC-Var	Bal. Acc
		FS → OS			
	SJ	50.08 ± 3.08	55.99 ± 4.74	49.55 ± 2.04	29.25 ± 2.04
	DF	45.46 ± 2.66	46.99 ± 1.98	49.79 ± 1.55	45.63 ± 4.53
	SM	35.08 ± 2.63	**67.09 ± 2.12**	50.47 ± 1.07	30.71 ± 0.67
	PV	**75.60 ± 6.16**	74.14 ± 4.23	49.80 ± 2.02	43.40 ± 5.09
	SD	65.72 ± 1.79	68.60 ± 1.70	46.65 ± 1.27	36.81 ± 0.74
	S4	59.01 ± 3.61	59.31 ± 3.59	48.31 ± 1.53	39.71 ± 5.70
		OS → FS			
	PV	45.05 ± 3.20	49.59 ± 2.34	50.20 ± 3.83	33.34 ± 0.16
	SM	**57.14 ± 2.16**	54.69 ± 1.45	48.15 ± 3.19	33.10 ± 0.40
	S4	49.28 ± 1.65	49.20 ± 1.68	50.11 ± 2.82	32.95 ± 0.18
	SD	58.08 ± 3.10	**58.01 ± 3.13**	48.81 ± 3.38	31.98 ± 4.07
	DF	51.29 ± 3.51	52.84 ± 2.65	49.03 ± 3.65	31.58 ± 0.63
	SJ	51.65 ± 2.31	52.64 ± 1.47	49.60 ± 2.75	30.46 ± 1.25

Table 10: Cross-dataset per-class accuracy. Direction-dependent class collapse evident: Tanker near 0% in OpenSARShip to FUSARShip; Fishing ≈ 0% in both directions. These per-class failures drive aggregate balanced accuracy collapse and motivate uncertainty-based error detection.

Model	Per-class Acc			Acc	BAcc
	Cargo	Tanker	Fishing		
		FUSARShip to **OpenSARShip**			
SJ	77.59 ± 9.41	5.73 ± 1.74	4.43 ± 4.08	58.15 ± 6.50	29.25 ± 2.04
DF	75.04 ± 6.55	2.28 ± 0.88	59.57 ± 16.59	56.48 ± 4.64	**45.63 ± 4.53**
SM	65.27 ± 4.78	26.58 ± 5.08	0.29 ± 0.97	54.31 ± 2.32	30.71 ± 0.67
PV	67.11 ± 9.09	4.10 ± 0.85	59.00 ± 21.95	51.13 ± 6.35	43.40 ± 5.09
SD	42.39 ± 2.97	68.04 ± 2.68	0.00 ± 0.00	48.01 ± 1.77	36.81 ± 0.74
S4	10.19 ± 1.37	85.79 ± 9.42	23.14 ± 25.60	29.42 ± 2.03	39.71 ± 5.70
		OpenSARShip to **FUSARShip**			
PV	**99.76 ± 0.29**	0.00 ± 0.00	0.26 ± 0.51	64.49 ± 0.32	**33.34 ± 0.16**
SM	99.29 ± 1.20	0.00 ± 0.00	0.00 ± 0.00	64.11 ± 0.77	33.10 ± 0.40
S4	98.85 ± 0.54	0.00 ± 0.00	0.00 ± 0.00	63.82 ± 0.37	32.95 ± 0.18
DF	93.85 ± 1.71	0.27 ± 1.31	0.61 ± 1.09	60.79 ± 1.08	31.58 ± 0.63
SJ	88.17 ± 4.73	3.20 ± 3.33	0.00 ± 0.00	57.11 ± 2.95	30.46 ± 1.25
SD	81.81 ± 6.42	14.13 ± 13.24	0.00 ± 0.00	53.63 ± 3.98	31.98 ± 4.07